Constructing the Face of Network Data

Ertza Warraich and Muhammad Shahbaz
Purdue University

CCS CONCEPTS

• Networks → Network measurement; **• Computing methodologies → Machine learning**;

KEYWORDS

Deep learning, GAN, network measurement, dataset generation

ACM Reference Format:
Ertza Warraich and Muhammad Shahbaz. 2021. Constructing the Face of Network Data. In *ACM SIGCOMM 2021 Conference (SIGCOMM '21 Demos and Posters), August 23–27, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3472716.3472852

1 INTRODUCTION

Network datasets are essential part of understanding, managing, and operating modern wide-area, data-center, and cellular networks. They are involved throughout the stages of network development, from simulations, stress testing, to machine-learning training (anomaly-based intrusion detection systems) and more. Despite the need, network datasets are rare due to concerns related to information privacy and sensitivity.

The datasets that we do have, are typically provisioned as feature sets (as described in [10]), instead of actual packet traces. Typically in a CSV format, the given dataset can help with machine-learning applications; however, without the payload, it cannot be used for running simulations and other forms of network testing. Their use in machine-learning approaches is even limited, as the feature-set is pre-selected and defined by the dataset provider; hence, making it entirely impossible to extract any new features that might be of interest.

Researchers have been working hard on overcoming the scarcity of such datasets [28]; however, the changing nature of network traffic makes whatever dataset they collect, quickly outdated. For example, it was thought that KDD99 [21] would be the final representative network dataset. However, a follow-up research [9], in the later years, argued that it is time to retire the KDD99 dataset with a new one. Other similar datasets [12, 18, 22] were presented and later disputed as being not representative of the current network traffic [1–3, 24, 25, 29].

In this paper, we aim to tackle this challenge and put forth a method, based on Generative Adversarial Networks (GANs) [13], for generating new (and timely) datasets, automatically, that are provisioned as complete raw packets traces of a network and not just feature values. GANs were initially used to generate image

SIGCOMM '21 Demos and Posters, August 23–27, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8629-6/21/08.
https://doi.org/10.1145/3472716.3472852

Feature	Description
Sport	Source Port
TotPkts	Total packets exchanged
TotBytes	Total bytes exchanged
SrcPkts	Source packets sent
SrcBytes	Source bytes sent
sTtl	Source time to live
dTtl	Destination time to live
SintPkt	Source inter-packet arrival time
DintPkt	Destination inter-packet arrival time
SrcWin	Source tcp window size
DstWin	Destination tcp window size

Table 1: List of network features from the CDX dataset used in our GAN framework.

datasets [32], but have since been used in a variety of domains, including text [8] and video [7, 31] generation to redesigning cities [14]. We take inspiration from these efforts and propose similar GAN-based system for generating network datasets.

The challenge here is to identify and extract key features from existing network traffic—similar to extracting primary features of a face in images—when constructing new traffic traces. Feature extraction is the most crucial step of any machine-learning system and directly impacts its overall effectiveness. A number of techniques and strategies for retrieving features have existed in the machine-learning literature for decades, and continue to evolve, e.g., (un)supervised neural-network learning [5, 26], as well as real-time [19, 30] and bayesian decision procedures [16]. We implement a similar GAN-based procedure but apply it to networking to identify what features make network datasets unique from other datasets, and what features account for variations in classes among the network dataset—we call this *constructing the face of network data.*

2 PROPOSED APPROACH

We implement an LSTM-based GAN [11] for our dataset-generation framework. LSTMs would work best in our case as the network traffic makes more sense when looked at as a collection of packets in flows instead of as individual packets. We choose Cyber Defense Exercise (CDX) [20] as a training input dataset, and perform feature extraction using ARGUS [6]. Furthermore, we filter the extracted features based on two metrics: (1) features that have highest classification accuracy, and (2) features that have a strong impact on the shape of the network traffic.

For the CDX dataset, we find eleven (11) features that fit these criteria (Table 1). We input these features to our GAN (which learns on them), and feeds new traffic features into our packet-generation tool, which outputs network traffic as *pcap* files with random data bytes as payload, rather than providing us with only the feature-value sets. Given an input trace (or dataset), the proposed framework can

Model	Classification Result	
	Skype	Other
Logistic Regression	100.00%	0.00%
Multi-layer Perceptron	100.00%	0.00%
K-Nearest Neighbors	99.04%	0.96%
Decision Tree	99.16%	0.84%
Random Forest	100.00%	0.00%
AdaBoost	100.00%	0.00%

Table 2: Classifying our framework's traffic.

generate any new (never-seen-before) network datasets—without exposing any critical and private details of the original dataset.[1]

To test the new dataset, we feed this data into a network using a pcap-based traffic generator (e.g., Scapy) to evaluate our framework's effectiveness. We implement a scenario, where a machine-learning model has to classify between Skype and other traffic and only allow the Skype traffic to pass through. We run six previously published network classifiers [17, 23, 27]: Logistic Regression, Multi-Layer Perceptron, K-Nearest Neighbor, Decision Tree, AdaBoost, and Random Forest. Each of them exhibits high classification accuracy when tailored to our application, correctly classifying normal traffic from Skype and vice versa.

Lastly, to construct the face of our network dataset, we extract all basic flow-based features and the statistical features (as used in DeltaShaper [4]). We extract and select a total of 256 distinct features, arrange them into a 16x16 array, and then normalize their values between 0–255. This essentially translates each input packet flow into a grayscale image representation. The resultant image is then fed to a StyleGAN [15], modified to work with our custom network image.

3 PRELIMINARY RESULTS

Table 2 shows the accuracy of the six classifiers when operating on the new datasets generated with our proposed GAN-based framework. The models were able to distinguish Skype traffic from other; hence, showing that the dataset was reflecting the true (non-)Skype traffic. One interesting outcome of this project is that, besides creating new datasets, the proposed framework can be used to morph existing traffic into some other traffic (e.g., Skype vs YouTube) for the purposes of circumventing censorship, covert communication channels, bypassing anomaly detection systems, and more.

The result of our StyleGAN is a grid of images shown in Figure 1, generated by GAN after learning and distinguishing the most important features of the input data. The grid contains a number of images, therefore we further extract singular 16x16 grayscale image from the grid. The extracted image is then enhanced and extrapolated to help visualize the face of network data. The resultant image is presented in Figure 2. Examining the pixel values and weights assigned to different input features, we see the following as core/face features of network data: 90th & 80th Percentile of PacketTimes, 90th & 80th Percentile of PacketTimesIn, skewPacketTimesIn, variancePacketTimesIn, skewPacketTimesOut, variancePacketTimesOut. Similarly, the least important features identified by our GAN are: InBurst Kurtosis, InBurst Skewness, InBurst Stdev, InBurst Variance, InBurst total.

[1]We plan to evaluate this further using more datasets, as future work.

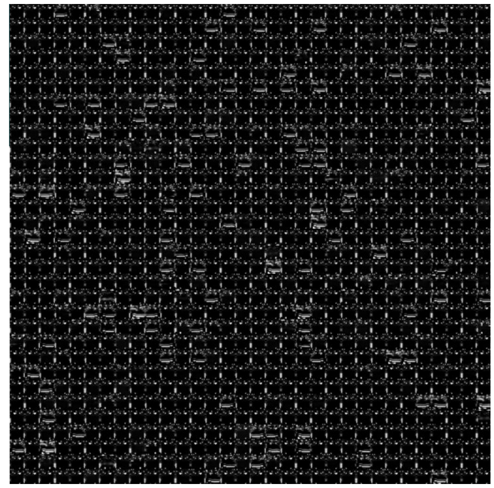

Figure 1: A grid of network data faces

Figure 2: Zoomed-in face of network data

4 CONCLUSION

We implement a GAN-based dataset-generation framework that can be used to create new network datasets. We implemented a new GAN to help identify the crucial features of network traffic. The visualization of network traffic in this manner brought an insight into what features make up its face, and what features define the variations in it while staying in the same class of data. Our evaluation concludes that inter-packet arrival times are the most crucial features while in-bursts are the least important in a CDX dataset.

In the future, we plan to work on classifiers that are able to distinguish between real data and GAN-generated data. We also plan to further evaluate the "face" features of network data to see how classification based on these *face* features might uncover new classes, which are not dependent on traffic type, but rather on parameters that affect the network performance, including usage, bandwidth, and jitter.

5 ACKNOWLEDGEMENTS

We thank the anonymous SIGCOMM reviewers for their invaluable feedback that helped strengthen the final draft. Our special thanks to the CS department at Purdue University for providing the opportunity and infrastructure to make this study possible.

REFERENCES

[1] Al-Hadhrami, Y., and Hussain, F. K. Real Time Dataset Generation Framework for Intrusion Detection Systems in IoT. *Future Generation Computer Systems 108* (2020), 414–423.

[2] Al-Kasassbeh, M., Al-Naymat, G., and Al-Hawari, E. Towards Generating Realistic SNMP-MIB Dataset for Network Anomaly Detection. *International Journal of Computer Science and Information Security 14*, 9 (2016), 1162.

[3] Alhaidari, F. A., and Alrehan, A. M. A Simulation Work for Generating a Novel Dataset to Detect Distributed Denial of Service Attacks on Vehicular Adhoc Network Systems. *International Journal of Distributed Sensor Networks 17*, 3 (2021), 15501477211000287.

[4] Barradas, D., Santos, N., and Rodrigues, L. E. DeltaShaper: Enabling Unobservable Censorship-resistant TCP Tunneling over Videoconferencing Streams. *Proceedings on Privacy Enhancing Technologies 2017*, 4 (2017), 5–22.

[5] Becker, S., and Plumbley, M. Unsupervised Neural Network Learning Procedures for Feature Extraction and Classification. *Applied Intelligence 6*, 3 (1996), 185–203.

[6] Bullard, C. ARGUS: The Network Audit Record Generation and Utilization System. http://www.qosient.com/argus. Accessed on 02/03/2021.

[7] Cai, H., Bai, C., Tai, Y.-W., and Tang, C.-K. Deep Video Generation, Prediction and Completion of Human Action Sequences. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018).

[8] Chen, L., Dai, S., Tao, C., Shen, D., Gan, Z., Zhang, H., Zhang, Y., and Carin, L. Adversarial Text Generation via Feature-Mover's Distance. *arXiv preprint arXiv:1809.06297* (2018).

[9] Creech, G., and Hu, J. Generation of a new ids test dataset: Time to retire the kdd collection. In *2013 IEEE Wireless Communications and Networking Conference (WCNC)* (2013).

[10] Damasevicius, R., Venckauskas, A., Grigaliunas, S., Toldinas, J., Morkevicius, N., Aleliunas, T., and Smuikys, P. LITNET-2020: An Annotated Real-world Network Flow Dataset for Network Intrusion Detection. *Electronics 9*, 5 (2020), 800.

[11] Gers, F. A., Schmidhuber, J., and Cummins, F. Learning to Forget: Continual Prediction with LSTM. *Neural computation 12*, 10 (2000), 2451–2471.

[12] Gogoi, P., Bhuyan, M. H., Bhattacharyya, D., and Kalita, J. K. Packet and Flow Based Network Intrusion Dataset. In *International Conference on Contemporary Computing* (2012).

[13] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative Adversarial Nets. In *Advances in neural information processing systems* (2014).

[14] Ibrahim, M. R., Haworth, J., and Christie, N. Re-designing Cities with Conditional Adversarial Networks. *arXiv preprint arXiv:2104.04013* (2021).

[15] Karras, T., Laine, S., and Aila, T. A Style-based Generator Architecture for Generative Adversarial Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019).

[16] Lowe, D., and Webb, A. R. Optimized Feature Extraction and the Bayes. *IEEE Transactions on Pattern Analysis and Machine Intelligence 13*, 4 (1991), 355.

[17] McGaughey, D., Semeniuk, T., Smith, R., and Knight, S. A Systematic Approach of Feature Selection for Encrypted Network Traffic Classification. In *2018 Annual IEEE International Systems Conference (SysCon)* (2018).

[18] Moustafa, N., and Slay, J. UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems (UNSW-NB15 Network Data Set). In *2015 military communications and information systems conference (MilCIS)* (2015).

[19] Nguyen, D., Memik, G., Memik, S. O., and Choudhary, A. Real-time Feature Extraction for High Speed Networks. In *International Conference on Field Programmable Logic and Applications* (2005).

[20] NSA. Cyber Research Center CDX Dataset. https://www.westpoint.edu/centers-and-research/cyber-research-center/data-sets. Accessed on 02/03/2021.

[21] Olusola, A. A., Oladele, A. S., and Abosede, D. O. Analysis of KDD'99 Intrusion Detection Dataset for Selection of Relevance Features. In *Proceedings of the world congress on engineering and computer science* (2010).

[22] Payer, G. Realistic Computer Network Simulation for Network Intrusion Detection Dataset Generation. In *Next-Generation Robotics II; and Machine Intelligence and Bio-inspired Computation: Theory and Applications IX* (2015).

[23] Perera, P., Tian, Y.-C., Fidge, C., and Kelly, W. A Comparison of Supervised Machine Learning Algorithms for Classification of Communications Network Traffic. In *International Conference on Neural Information Processing* (2017).

[24] Pham, V. C., Makino, Y., Pho, K., Lim, Y., and Tan, Y. IoT Area Network Simulator For Network Dataset Generation. *Journal of Information Processing 28* (2020), 668–678.

[25] Sangster, B., O'Connor, T., Cook, T., Fanelli, R., Dean, E., Morrell, C., and Conti, G. J. Toward Instrumenting Network Warfare Competitions to Generate Labeled Datasets. In *CSET* (2009).

[26] Setiono, R., and Liu, H. Feature Extraction via Neural Networks. In *Feature Extraction, Construction and Selection*. Springer, 1998, pp. 191–204.

[27] Shahraki, A., Abbasi, M., and Haugen, Ø. Boosting Algorithms for Network Intrusion Detection: A Comparative Evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost. *Engineering Applications of Artificial Intelligence 94* (2020), 103770.

[28] Thakkar, A., and Lohiya, R. A Review of the Advancement in Intrusion Detection Datasets. *Procedia Computer Science 167* (2020), 636–645.

[29] Vasudevan, A., Harshini, E., and Selvakumar, S. SSENet-2011: A Network Intrusion Detection System Dataset and its Comparison with KDD CUP 99 Dataset. In *2011 second asian himalayas international conference on internet (AH-ICI)* (2011).

[30] Wu, F., Jiang, X., Ma, W., Wang, L., Jiang, Y., Guan, S., Li, X., Song, M., Liu, M., and Yin, M. A Feature Extraction Method of Network Traffic for Time-Frequency Synchronization Applications. In *International Conference on Computer Systems, Electronics and Control (ICCSEC)* (2017).

[31] Yan, W., Zhang, Y., Abbeel, P., and Srinivas, A. VideoGPT: Video Generation using VQ-VAE and Transformers. *arXiv preprint arXiv:2104.10157* (2021).

[32] Yang, J., Kannan, A., Batra, D., and Parikh, D. LR-GAN: Layered Recursive Generative Adversarial Networks for Image Generation. *arXiv preprint arXiv:1703.01560* (2017).